

School of Computer Science

Machine Learning in Fulfilment of

SPEC9270

Maksymilian Drzezdzon

C15311966

Degree: TU060/1

Module Coordinator: Sarah Jane Delany

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged: 🗹

**Date: 2021/03/21**

Acknowledgments

Collection of this dataset was made at the Mind Research Network under an NIH NIGMS Centers of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 to Vince Calhoun (PI).

Definition of problem – Kaggle Contest Entered

The Kaggle contest entered is the **MLSP 2014 Schizophrenia Classification Challenge** (Network, 2014), although the competition is over it is still possible to submit work and receive an evaluation. My reasoning for picking this competition is that I wanted to study psychology and go into research particularly in dissociative disorders and schizophrenia, however due to personal circumstances at the time I went with engineering.

Today, I’m still interested in a PhD and have gained a different perspective I otherwise wouldn’t have had, one that sees that there are a lot of holes and short comings in traditional diagnosis of mental disorders such as D.I.D and schizophrenia. All of which hopefully can be dealt with machine learning (ML).

Some of these short comings are that there is no process to date that properly diagnoses D.I.D, despite it being acknowledged as a mental illness in the 1950s, research between then and now has been done but was later found to be fraudulent or difficult to reproduce. (David Blihar, 2020)

While remaining brief and acknowledging other issues that stem from the use of machine learning (ML), MRI and fMRI images can be fed to ML algorithms in order to be used as a tool that can help clinicians avoid misdiagnosis. Young girls tend to not be diagnosed with ADHD because it manifests differently (CONNOLLY, 2020), having a better system for such differences in mental health would help avoid undiagnosed patients in situations such as school performance.

Currently the best/popular go to method is MRI image classification as they are interpretable by clinicians and can be examined after a ML model points to one claiming “that one there seems off”, this is temporary solution to deep learning techniques being powerful for classification yet uninterpretable which is a huge deficit for application in healthcare. This is why, currently, I think ML will be another tool for diagnosis not a replacement or augmentation of existing methods.

Having said that, there are other areas such as using NLP for diagnosing/detecting the onset of Alzheimer’s disease (Elif Eyigoza, 2020).

This is where I want to step in, I want to spend time to develop skills and knowledge in machine learning, leveraging my engineering experience in order to transition into ML in mental health, an area my workplace (IBM) is active in.

This project hopefully will be an introduction to that, the dataset used in this project is in a different format then what I expected, I approach it as a learning opportunity for the future.

Models Built – Model Configuration

**Data preparation**

There was very little required to prepare data for this assignment. Originally, I was under the impression that the data consisted of MRI images, however it’s a set of correlation values and xxx

There are no missing or incomplete records in this dataset otherwise they would need to be removed or filled in if possible. There is an option to use a scalar to reduce the gaps between values however these values are already very small and close together.

The use of visualisations could be used to see the distribution of data and a boxplot to get a five number summary for a quick evaluation.

One way of working with MRI images when there is a shortage of data is to rotate them a few times to synthesis more data points for a machine learning algorithm to pick up on.

List other ways you can prepare data here

**Implementation**

Models used in this report are Support vector machine, as it was very popular in research papers I’ve read cited in at the beginning of this paper, most of which scoped the application of machine learning methods to mental health diagnosis.

**SVM/SVC**

SVM uses something called a support vector classifier in machine learning which has worked the best without any hyperparameter tuning and only improved once a better kernel was adopted. SVM preforms very well even when there is limited data available, this is probably why it’s used for mental illness classification.

There’s usually very limited data with potential for a lot of noise. That noise being overlapping conditions that are known to coexist with certain disorders like D.I.D or schizophrenia. This is further supported when comparing results for public and private leader boards, there is only a 2.2% difference between scores, yet, the public leader board uses 1% of data.

SVM uses a hyperplane to divide data points, a hyperplane is a line that measures the distance between all points, this distance is referred to as a margin, for multimodal data SVM will add another dimension to better separate groups. To my understanding this is why it preformed so well without any optimisation.

This model was trained using the following configs:

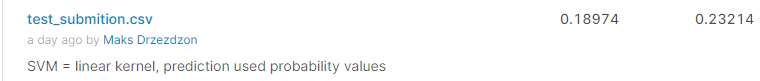
X\_train, X\_test, y\_train, y\_test = train\_test\_split (data, labels, test\_size=0.3, random\_state=109)

clf = svm.SVC(kernel='sigmoid') # Best Kernel was Sigmoid

explain params used in svm.SVC https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Here data was split in a ratio of 3:7 test/training, random state is variable used how data will be randomly mixed/split, choosing a number will make it consistent, 109 produced the best result, if this variable is not set to a number model results will be different each time.

When creating machine learning models for healthcare its rather difficult to interpret a probability value, for instance if a patient has cancer and its probability of being malignant is 40%. The sigmoid kernel returns a binary value similarly to a logistic regression. https://dataaspirant.com/svm-kernels/

When trying the linear kernel and with the probability function 

Its very clear that binary values are expected.

The sigmoid kernel preformed best, to my limited understanding this is because it adds non-linearity when choosing which values to pass as an output and which one not to pass, if a result is > or equal to 0.51 it will round up to a 1 and likewise for 0 if its below 0.51. (Kumawat, 2019) It yielded



Random Forest

<https://datascience.stackexchange.com/questions/6838/when-to-use-random-forest-over-svm-and-vice-versa>

Neural Network

Local Evaluation – Evaluation Strategy

When comparing models and reading what each leader board score is for, I found that the private leader board score is the one that uses more data and is more relevant for the actual performance of the model. (Park, 2012).

Model evaluation consisted of:

Accuracy 🡪 fraction of predictions the model got right.

Precision 🡪 Refers is a ration of correctly predicted positives to the total of predicted positive observations.

Recall 🡪 Meanwhile recall refers to the number of those instances that are correctly predicted.

Cohen’s Kappa 🡪 This is a metric used to assess percent agreement, this is a useful metric because it also takes into account the possibility of said agreement occurring by chance. (McHugh, 2012) This is accomplished by taking the difference between the overall accuracy and the overall accuracy that can be obtained by chance. This metric is used when working with unbalanced datasets. Its an informative way to if the dataset is unbalanced or not as unbalanced datasets will produce a lower value.

Log Loss/Cross-Entropy 🡪 measures the performance of a classification model, however it requires probability values rather than binary values, for the current setup this metric was not used buts its worth mentioning as it seems popular and useful for models where probability values can be used.

Learning Curve 🡪 This ‘metric’ or more so visualisation maps the change in performance as more values are added.

ROC 🡪 This is a graph that shows the performance of a classification model at all thresholds

AUC 🡪 This is a graph that visualises the are under the ROC curve, this can be interpreted as the probability that the model will rank a random positive example more highly than a negative example. It measures how well a prediction is ranked. (TAVISH SRIVASTAVA, 2019)

Usually for binary classification, precision, recall and F1 score should suffice. More metrics were used to get a better understanding of their application in classification models and they provide a better overview of the model itself.

K-fold Cross-Validation 🡪 data is partitioned into k number of subsamples; this is repeated k number of times and each time one of the subsets is used as the set/validation set meanwhile the k-1 subset is used to train the model.

Local Evaluation – Evaluation Results

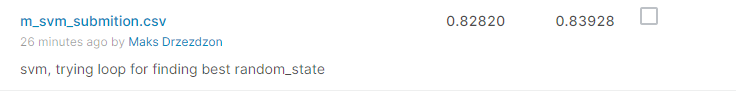
Kaggle – Performance Report

This competition closed 7 years ago, with 313 participants.

Best attempt with SVM:

Public: 220/313

Private: 217/313



To-do’s – Further work

More time invested in understanding other methods of assessing model performances, similar to Cohens Kappa. Investigating other methods for handling multimodal data, along with allocating more time for feature selection. There is a lot of space for improvement which is something I’d how to achieve in my thesis using this dataset. Some methods to research later would be distance weighted discrimination function, other ways of using SVM, this can be accomplished by reading more research papers. Trying multi-layer neural networks using the previously used sigmoid function as an activation method.

# References

CONNOLLY, M. (2020, Jan 17). *adhdireland*. Retrieved from Easy-to-Miss ADHD Symptoms in Girls: https://adhdireland.ie/what-adhd-looks-like-in-girls/

David Blihar, E. D. (2020). A systematic review of the neuroanatomy of dissociative identity disorder. *Elsevier*, https://doi.org/10.1016/j.ejtd.2020.100148.

Elif Eyigoza, S. M. (2020). Linguistic markers predict onset of Alzheimer's disease. *Elsevier*, 10.1016/j.eclinm.2020.100583.

Kumawat, D. (2019, August 22). *analyticssteps*. Retrieved from 7 Types of Activation Functions in Neural Network: https://www.analyticssteps.com/blogs/7-types-activation-functions-neural-network

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276–282.

Network, M. R. (2014). *MLSP 2014 Schizophrenia Classification Challenge - Diagnose schizophrenia using multimodal features from MRI scans*. Retrieved from Kaggle: https://www.kaggle.com/c/mlsp-2014-mri

Park, G. (2012, July 6). *gregpark.io/blog/Kaggle-Psychopathy-Postmortem*. Retrieved from The dangers of overfitting: a Kaggle postmortem: http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/#:~:text=The%20public%20leaderboard%20scores%20are,subset%20of%20the%20test%20data.&text=Top%20of%20the%20private%20leaderboard,after%20the%20contest%20is%20closed.

TAVISH SRIVASTAVA. (2019, August 6). *11 Important Model Evaluation Metrics for Machine Learning Everyone should know*. Retrieved from analyticsvidhya.com/: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/